

Conversations in the Wild: Data Collection, Automatic Generation and Evaluation

Nimra Zaheer^a, Agha Ali Raza^b, Mudassir Shabbir^a

^a*Computer Science Department, Information Technology University, Lahore, Pakistan,*

^b*Computer Science Department, Lahore University of Management Sciences, Lahore, Pakistan,*

Abstract

The aim of conversational speech processing is to analyze human conversations in natural settings. It finds numerous applications in personality traits identification, speech therapy, speaker identification and verification, speech emotion detection, and speaker diarization. However, large-scale annotated datasets required for feature extraction and conversational model training only exist for a handful of languages (e.g. English, Mandarin, and French) as the gathering, cleaning, and annotation of such datasets is tedious, time-consuming, and expensive. We propose two scalable, language-agnostic algorithms for automatically generating multi-speaker, variable-length, spontaneous conversations. These algorithms synthesize conversations using existing non-conversational speech datasets. We also contribute the resulting datasets (283 hours, 50 speakers). As a comparison, we also gathered the first spontaneous conversational dataset for Urdu (24 hours, 212 speakers) from public talk shows. Using speaker diarization as an example, we evaluate our datasets and report the first baseline diarization error rates (DER) for Urdu (25% for synthetic dataset-based models, and 29% for natural conversations). Our conversational speech generation technique allows training speaker diarization pipelines without the need for preparing huge conversational repositories.

Keywords: conversations, conversational datasets, dataset generation, corpus generation, synthetic dataset generation, speaker diarization, language-agnostic, Urdu.

1. Introduction

Humans are social beings having innate abilities to interact. In addition to gestures and expressions, a verbal conversation is a fundamental element in human communication. Beyond pure linguistic content, conversations also contain personality traits of individuals including conscientiousness [1], empathy [2], and emotional cues [3]. To analyze such traits, numerous conversational speech datasets are available albeit, for a handful of languages such as English [4, 5, 6], French [7] and Mandarin [8]. With 171 million speakers, Urdu is the 11th most widely spoken language in the world [9]. Despite this, we do not find any conversational datasets or conversational speech processing applications available for Urdu. However, we do find several non-conversational datasets of Urdu spontaneous speech. The Baang dataset contains 1, 207 hours of telephonic speech extracted from public forums incorporating 4, 678 speakers [10]. This speech repository has asynchronous audio posts of various users of a social media platform hence, real-time multi-user one-to-one correspondence, turn-taking, and speech overlaps are not captured in these utterances. Another repository developed for speech activity detection (SAD) contains 3000 hours of telephonic conversations for Urdu and other languages [11]. This repository is not multi-speaker, and the annotation is restricted to speech activity detection which is just one module inside the speaker diarization pipeline. These datasets cannot be used for multi-speaker conversational speech processing. Further, these repositories contain purchasable licenses and are not publicly available to the research community. A dire need exists for a natural conversational speech database that could be utilized for analyzing traits of conversations and their participants.

In this paper, we provide a scalable approach to generate language-agnostic datasets by proposing two algorithms: *Autogen1* and *Autogen2* that automatically generate spontaneous multiparty conversations along with ground labels from unique speaker profiles that mimic the real-time natural human discussions embedded with diverse ambient noise profiles. Additionally, we provide individual speaker profiles and spontaneous conversations: *Autogen1 database* and *Autogen2 database* generated from the aforementioned algorithms. We also discuss the development of the first 24- hours dataset: *Conversational speech repository for Urdu Language (CONVURL)* that contains natural spontaneous conversations in Urdu with 212 unique speakers. Each clip contains at least five speakers (with possible speech overlaps) fo

cusing on scholarly debates and political talk shows. Furthermore, each clip has been manually annotated and provided with standardized NIST Rich Transcription Time Marked (RTTM), and Unpartitioned Evaluation Map (UEM) files [12].

Conversational datasets can be evaluated using numerous applications of which speaker diarization is an important one – the question of *who spoke and when*. Diarization requires large conversational speech repositories. Recent works utilize deep learning algorithms to reduce speaker diarization error rate (DER) using a robust open-source toolkit named *Pyannote Audio* [13]. Such systems provide speaker diarization baseline results for English. We use speaker diarization to perform extensive experimentation and report that both natural and synthetic conversations aid speaker diarization problems. To the best of our knowledge, these are the first Urdu speaker diarization benchmarks. Moreover, we show that synthetic conversations can be utilized for training speaker diarization pipelines by using transfer learning without the need for the preparation of manually annotated datasets. We support this claim by performing experiments on AMI and CONVURL corpus. In summary, this work makes the following contributions:

1. We present two language-agnostic *algorithms* for automatically generating synthetic multi-speaker conversations from non-conversational datasets.
2. We develop the first multi-speaker conversational speech repository for Urdu with standard annotations.
3. We provide the first baseline results for Urdu speaker diarization using the aforementioned datasets.
4. Lastly, we report novel benchmarks based on fine-tuning pre-trained speaker diarization models (English-Urdu, Urdu-English).

2. Related Work

Some of the widely used datasets involving conversations include CALLHOME (LDC2000IS97), AMI [4], ICSI [5], and CHIME [6]. CALLHOME or NIST SRE 2000 (Disk-8) is a 20-hour long repository containing 2 to 7 speakers engaged in telephonic communication in six languages including Arabic, English, German, Japanese, Mandarin, and Spanish. AMI corpus is a 100-hour dataset targeting meeting sessions conducted by 3 to 5 speakers in English [4]. 72-hour ICSI meeting corpus caters increased number of

speakers ranging from 3 to 10 engaged in meeting sessions of four different types [5]. CHIME-5 contains 50 hours of speech containing four speakers recorded in the kitchen and living rooms of a house [6]. VoxConverse is a recent dataset containing conversations retained from Youtube videos with a maximum of 21 speakers per conversation [14]. Apart from CALLHOME, all datasets are in English. A set of DIHARD challenges were also introduced for robust speaker diarization [15].

There also exists an audio-visual conversational speech repository containing 30 minutes long conversation with 120 dialogues spoken in the Swedish [16]. E-Tape is a 30-hour long repository consisting of radio and TV broadcasts covering multiple speaking styles available for the French [7]. A 200-hour telephonic speech corpus is available for Mandarin containing 2100 speakers. There are 1206 audio clips of 10-minute duration containing spontaneous conversations [8].

Conversational databases for low-resource languages such as Hindi and Turkish also exist for speaker recognition [17, 18]. Such datasets for low-resource languages are hard to find and restricted to the speaker recognition task. Considering natural spontaneous speech conversations, only a handful of such datasets exist for Urdu. One such repository contains 1,207 hours of telephonic speech extracted from open public forums incorporating 4,678 speakers [10]. This speech repository has asynchronously recorded audio posts of various users hence real-time, multi-user, one-to-one correspondence, and speech overlaps are not captured in such utterances. Another dataset developed for speech activity detection contains 3,000 hours of telephone conversations for Urdu and other languages [11]. This repository is void of multi-speakers and annotation is restricted for speech activity detection which is just one module inside the speaker diarization pipeline. Another 45 hours corpus incorporating 82 speakers exists for speaker-independent automatic speech recognition system [19]. Such datasets cannot be used for multi-speaker diarization systems because of the ground labels being limited to a particular module in the diarization pipeline and a lack of real-time spontaneous speech. Further, these repositories contain purchasable licenses or are publicly not available to the research community. There is a dire for a natural conversational speech database that could be utilized for discerning interesting insights into conversations and their participants. In this paper, we develop the first 24- hours dataset: CONVURL which contains natural spontaneous conversations in Urdu with 212 unique speakers. Table 1 enlists the conversational datasets with descriptions.

All of the former repositories have been carefully collected, prepared, and annotated manually and made available for further experimentation. Such datasets require a lot of time, effort, and resources for collection and preparation. Also, these repositories exist only for a few languages whereas nearly 7,000 languages are being spoken all over the world [9]. Further, to automate such a process, an algorithm must be devised to generate conversations mimicking real-time conversations. Recent research is focused on improving dialogue generation systems for human-robot conversations which produce responses to spoken conversations. Most of this work focuses on Seq2seq models which produced incoherent responses [20]. Improvements were made by using ensemble technique [21], utilizing additional knowledge [22], reward function [23], and goal sequence planning [24]. While this work is much more suitable for machine-to-human interaction, it lacks realism in multi-party conversations taking place between humans. Datasets collected for speaker diarization are also available [25, 26, 27].

It would be interesting to predict speaking turns, personality traits, and emotional cues present in a conversation. To the best of our knowledge, there exists no technique which provides language-agnostic spontaneous conversation generation with variable time duration and the number of speakers. We devise a novel algorithm that automatically generates conversations with ground labels from a smaller dataset saving us from the laborious task of collection, preparation, and annotation. We are interested in making use of such datasets in the speaker diarization problem which deals with *who spoke and when*.

3. Methodology

Procuring a multi-speaker, multi-dialect, and multi-accent speech corpus for a language is an arduous task. Our goal is to provide an efficient algorithm for generating natural conversations without the need for extensive data collection and a laborious resource labeling process. Further, we prepare a natural conversational repository leveraging spontaneous talks present in the wild. This dataset can be utilized for the majority of applications involving conversations and also for evaluating our synthetic dataset. We describe the collection process for spontaneous natural conversations. Secondly, we discuss the process for artificially generating conversations followed by results.

Table 1: Summary of conversational datasets for languages (Lang.) English (En.), Urdu (Ur.), French (Fr.), Mandarin (Ch.), and Swedish (Sw.) along with the number of speakers (Sp.) per conversation.

Dataset	Lang.	Size (hr)	#sp.
CALLHOME	En.	20	2 – 7
AMI [4]	En.	100	3 – 5
ICSI [5]	En.	72	3 – 10
CHIME [6]	En.	50	4
DIHARD II Track 1,2 [15]	En.	46	1 – 8
DIHARD II Track 3,4 [15]	En.	293	4
VoxConverse [14]	En.	74	1-21
E-Tape [7]	Fr.	30	-
Spontal [16]	Sw.	60	-
HKUST/MTS [8]	Ch.	200	2
Baang Dataset [10]	Ur. + others	1,207	2
RATS [11]	Ur. + others	3,000	-
CONVURL	Ur.	24	4 – 5
Autogen 1 & 2	Ur.	259	5

3.1. Natural conversational speech dataset

The first and foremost step is to identify interesting discussion scenarios containing multiple speakers engaged in natural discussions such as meetings, lunch and dinner time discussions, press conferences, talk shows, courtroom decisions, online video lectures, scholar debates, and political talk shows. The aforementioned keywords were searched on Youtube which resulted in English videos. Thus, the search was reduced to selecting the videos using Urdu. These outcomes were carefully analyzed based on spoken language, number of videos, and speakers. We shortlisted religious scholar debates and political talks for our dataset: CONVURL. Numerous YouTube videos were screened to ensure that Urdu is spoken for the maximum time throughout the audio clip. Lastly, the rationale for choosing Urdu instead of code-mixed language was to perform language-agnostic experimentation, which will appear in the latter sections.

To mimic real-time human conversations, we carefully hand-pick videos containing discussions targeting at least five speakers with possible speech overlaps. Further, videos with a focal person such as an anchor or moderator with at least four participants were selected for meaningful discussions.

Scholarly debates and political talk shows in Urdu were finalized as the categories of choice because of spoken Urdu proficiency throughout the videos and the consistent structure of discussions in terms of the number of speakers. There are 18 clips spanning 12 hours and 17 minutes related to scholar debates where male moderators start a debate on a topic chosen from the religious domain and four participants are invited for discussions. In 18 of these audio clips, two moderators overlap throughout resulting in 99 unique speakers. Similarly, there are 20 clips worth 11 hours and 58 minutes containing political talks where female anchors initiate discussions among politicians focusing on national topics. Two anchors overlap in these videos resulting in a total of 113 unique speakers. The clips contain natural ambient noise along with starting, ending, and intermission music clips. Additional speakers are added by the anchors to the discussion depending on the scenarios. Also, no noise cancellation filters were applied. All clips have one channel (mono), 8-kHz sample rate, 16-bit sample size, and 128-kbps average bit rate. CONVURL is available to the research community for experimentation and possible extensions at <https://tinyurl.com/2p9x9ptb> under a Creative Commons Attribution-NonCommercial-ShareAlike license. The details for each category for the repository have been explained in Table 2.

3.1.1. Groundtruth collection and processing

Once the videos were obtained, the ground truth files had to be prepared for measuring the performance of the applications for conversational scenarios. The task for attaining the tags for each video is complex in nature [28, 29, 30]. Moreover, all supervised machine and learning algorithms utilize these datasets for training purposes, hence numerous user-friendly annotation tools for speech processing are available such as audino, PRAAT, Gecko, etc. [29, 31, 32].

For ground-truth preparation, each video has been manually annotated by introducing speaker tags with timestamps using a state-of-the-art annotation tool Gecko [32]. For a 40 minutes audio clip, it takes around three to four hours for annotation with full concentration as the speaker identities and speech overlaps have to be carefully marked. The tool provides a user-friendly interface for identifying and tagging the speaking time of each participant. Users can select a part of the audio clip and label the speaker with pre-defined identities. The ground-truth files are saved in the standardized NIST RTTM formats [12]. Furthermore, UEM files containing the boundaries of each file’s starting and ending duration are also provided with

Table 2: Statistics for CONVURL comprising of discussion category, number of clips, number of unique speakers (Spks.), average duration (Av. du.) per clip in minutes, and the total duration (Tot. dur.).

Cat.	No. of clips	Spks.	Av. du.	Tot. dur.
Scholarly Debates	18	99	40.9	12 h 17 m
Political Talk shows	20	113	35.9	11 h 58 m
Full repository	38	212	38.3	24 h 15 m

convurl.pdf

Figure 1: This figure elaborates the preparation of CONVURL starting from selection of conversational scenarios followed by retention of selected videos. Last step is to annotate each conversation to finally prepare the dataset.

audio clips. Figure 1 elaborates our data collection process for CONVURL. Due to the unavailability of resources, only one annotator was engaged to perform the task hence, a candidate with good Urdu speaking skills and holding a Master’s degree from Lahore city, as hometown, was selected to perform the annotation. It took approximately 100 hours for the annotator to tag a 24-hours dataset which is still a relatively small repository compared to state of the art. Therefore, there is a need to develop an automatic mechanism to generate natural conversations along with labels. Such a technique would produce training and testing data in a scalable manner with variables to control the number of speakers and duration of the audio clips. Furthermore, 10% of CONVURL has been annotated by another evaluator belonging to Lahore city, holding an intermediate education. It took 12-hours for the second evaluator to annotate the subset of these conversations. We have leveraged **Pygamma agreement** [33], an open-source package to measure the inter- evaluator’s score. This package takes RTTM files of different annotators and calculates the gamma- agreement. The value obtained resides between 0 and 1, closer to 1 means the evaluators’ annotations are similar. On 10% of CONVURL, we achieve 0.75, which is a satisfactory value.


3.2. Synthetic Conversational speech dataset

The structure of discussions plays an important part while generating synthetic conversations. While collecting CONVURL, it was observed that speakers talk with each other for variable time duration while changing attention from one speaker to another along with possible overlaps and pauses. Further, every speaker participating in the discussion has an ambient noise different from his/her fellow speakers due to factors like distance from the microphone (video/audio conferencing in a meeting room), fan positions, speaker orientation, the direction of arrival of sound on the microphone and different microphones (video conferencing). The difference in distance from the mic to the speaker also changes the loudness level within a discussion. Lastly, the number of speakers per conversation and the duration of the call vary according to the scenario. All the aforementioned factors were considered while generating synthetic conversations.

Apart from conversational settings, preparing ground labels for each discussion is a crucial and time taking step, which needs to be automated to reduce efforts and increase efficiency. As previously mentioned, ground labels are allotted to each person with their timestamp and duration of talking. This entire process is hard and the annotator is required to view the speakers from the video to assign speaker identities accordingly. Keeping in mind the aforementioned settings, we designed an algorithm for generating variable-length conversations.

We break down each conversation into two parts: *dialogue* where a speaker talks followed by *silence* where a gap is observed between two adjoining speakers. Moreover, the standard ground-truth file is tagged with the starting time and the duration of speech followed by the speaker's identity and some other formal tags which can be handled through the code. Formally, the ground truth file has 10 fields for one speaker turn [12]. Hence, if we can control the speakers and their talking, we can automatically generate conversations along with the gold labels.

To simulate a conversation, the first step is to collect audio clips with unique speaker profiles where only one speaker is talking throughout the clip. The next step is to prune these clips on silences and keep these smaller chunks against each speaker, this step will result in *dialogues* for each speaker. Further, we retrieve the silence profiles containing ambient noise for each clip which will serve as *silences*. We randomly choose the speakers from the list, retrieve speaker-wise chunks in a random order (without replacement), and finally fetch silence profiles randomly. Note that no dialogue will be repeated



autogen_generic.pdf

Figure 2: This figure elaborates on the process for automatic conversation generation. The first step is to extract unique speaker profiles followed by a cleaning step to eliminate any extra speakers. Further, clips are pruned on silences, and dialogue and noise profiles are sequenced together to form conversations.

since we selected clips in a random order without replacement. The next step is to pick one chunk from the shuffled array and a silence profile, place them in sequence and update the ground-truth file. We repeat this process till the required length of conversation is reached. Figure 2 elaborates on the summarized process of automatic conversation generation.

Next, we describe two algorithms for generating conversations. Automatically generated conversations will be denoted by *Autogen* followed by a number indicating variations.

3.2.1. Speaker profiles

An audio clip spoken by a unique speaker is defined as a speaker profile $s_i \in S$, where S is the set of all speakers. For the *dialogue* module, we have collected 50 unique speaker profiles of variable length containing 20 female and 30 male public representatives respectively. Each audio clip contains a speech for one unique speaker. For acquiring these clips, keywords like *<politician name> national assembly address* were queried on YouTube’s search engine. The clips where only the selected politician spoke were selected. Also, there were some clips embedded with starting and ending with another speaker, all such clips were trimmed so that only one speaker per audio file was retained. The rationale for selecting the keyword *address* was to ensure such clips where there was no communication with the media or other parties. After cleaning, we were left with 12 hours of speech. Lastly, to get dialogues of variable size, we segmented the whole clip based on silences. These speaker profiles along with the names of public representatives are available to the research community for experimentation and possible extensions at <https://tinyurl.com/2p9x9ptb> under a Creative Commons

Attribution-NonCommercial-ShareAlike license.

3.2.2. *Autogen*

We elaborate our various conversational settings in this section. Algorithm 1 represents the generic pseudo-code for our two variations namely *Autogen1* and *Autogen2*. The differences in the code are reflected in the colors: orange and teal for *Autogen1* and *Autogen2* respectively. In the *Autogen1* setting, unique speaker profiles are segmented into *dialogues* based on *silences* and kept in unique speaker folders. Further, silences in between dialogues greater than α second with the threshold of β decibels relative to full scale (dBFS) are extracted using the Pydub library [34]. These silences are maintained in one folder only, without any speaker identity because of the random selection of dialogues from the speakers’ folders. This step has been explained in Algorithm 1, line 3 – 7 and takes $\mathcal{O}(|S|)$ where S is the set of speaker profiles. The order will not matter as explained in the later steps.

We define the number of variables for generating conversations of the user’s choice. We denote the number of speakers per conversation by n , the number of conversations (outer bound) as m , and the maximum duration of conversation t (minutes). We randomly choose n speaker identities without replacement for each conversation. For each sequence, we retrieve all chunks against speaker identities and shuffle them, letting the size of this list be k . Next, we randomly select k silence profiles. Select speaker chunk and silence profile and embed them in sequence. Maintain the RTTM file for ground truth by appending the speaker id, start time, and total speaking time duration. Repeat this process till the duration of the conversation is less than or equal to t . Repeat the process for m conversations. there could be a case where dialogues and silences remain after making a conversation of t duration. We choose $\alpha = 0.5$ and $\beta = -40$ for each speaker profile. We also apply loudness normalization to each conversation.

As an example, for $n = 3, m = 2, t = 2$, Assume that speaker identities are selected in order: $[(1, 5, 8), (6, 9, 10)]$ indicating that the first conversation will contain speakers with identities 1, 5 and 8 and so on. The algorithm will collectively retrieve all chunks of speakers 1, 5 and 8 of size k . Also randomly retrieve k silence profiles. For m times, sequentially it will extract chunks and silence profiles to generate conversations $\leq t$. Repeat this for a sequence $(6, 9, 10)$. Algorithm 1 elaborates the steps involved in the construction of the conversation using *Autogen1*.

It is interesting to note that *Autogen1* has silence clips randomly embed-

ded throughout the conversation. The speaker’s noise profile will be concatenated with another speaker’s dialogue and noise profiles independent of speakers i.e. not present in the speaker profiles do not exist in these conversations. Thus, we propose another variation of the algorithm to generate conversations namely *Autogen2* which contains speaker dialogues embedded with their noise profiles and silences which are not present in any speaker’s clip.

For *Autogen2*, we extract γ variable-length ambient noise profiles from audio clips different from 50 speaker profile clips (Algorithm 1, line 14). Also, while chunking, we sequentially retrieve silence profiles and embed each chunk with this noise by splitting silence into two equal parts and adding them before and after each chunk in a prefix/postfix manner. This step has been explained in Algorithm 1, line 7. For the *silence* module, we randomly select noise from 20 clips and embed it in between the *dialogues* as mentioned in the former paragraph. Algorithm 1 elaborates the steps involved in the construction of the conversation using *Autogen2*. We set $\gamma = 20$ for the ambient noise profiles. Figure 3 elaborates the detailed process for *Autogen1* and *Autogen2* with $n = 2$. Algorithm 1 executes in $\mathcal{O}(|S| + m \times \sum_{j=1}^n \Theta_j)$, where $\Theta_j =$ is the total number of dialogues and silences for the speaker $s_j \in S$. Independent of speakers’ audio clips specifications, all conversations generated from the algorithms have one channel (mono), 352 kbps bit rate, and 22050-Hz sample rate.

4. Results

The following sections elaborate on the evaluation of our natural and synthetic datasets. We utilize these conversations in the speaker diarization problem by using an open-source pipeline *Pyannote.audio* [13]. Lastly, we perform extensive experimentation on *Convurl*, *Autogen1*, *Autogen2*, and *AMI* corpus to match the state-of-the-art results.

4.1. Speaker Diarization

The task of determining *who spoke when* in an audio signal is called speaker diarization. This problem has diverse applications e.g. meeting transcriptions, behavioral analysis, and automatic speech recognition systems [35]. The performance of this problem is measured in Diarization Error Rate (DER) which is composed of a sum of false alarm (duration of non-speech

Algorithm 1 Autogen

Input: Set of speaker profiles (S), number of speakers (n , $n \leq |S|$), number of conversations (m), duration of conversation (t), ambient noise $\notin S$ ($silence'$)

Output: saved conversations, RTTM, UEM, and LST files

```
1:  $silence \leftarrow \emptyset$ 
2:  $dialogues \leftarrow \emptyset$  : a dictionary for maintaining dialogues against each speaker
3: for  $s_i \in S$  do
4:   Autogen1:
5:    $silence, dialogues[s_i] \leftarrow$  Get  $s_i$ 's dialogues and silences from its respective speaker profile
6:   Autogen2:
7:    $dialogues[s_i] \leftarrow$  Embed silence in pre and post-positions of  $s_i$ 's dialogues
8: for  $i \leq m$  do
9:    $C \leftarrow$  extract  $dialogues$  of randomly selected  $n$  speakers without replacement
10:   $C \leftarrow$  randomly shuffle  $C$ 
11:  Autogen1:
12:   $Sil \leftarrow$  randomly choose  $|C|$  silences from  $silence$ 
13:  Autogen2:
14:   $Sil \leftarrow$  randomly choose  $|C|$  silences from  $silence'$ 
15:  Iterate over  $C$  and  $Sil$  to make conversations within time limit  $t$ 
16:  Update RTTM, LST, and UEM files
```

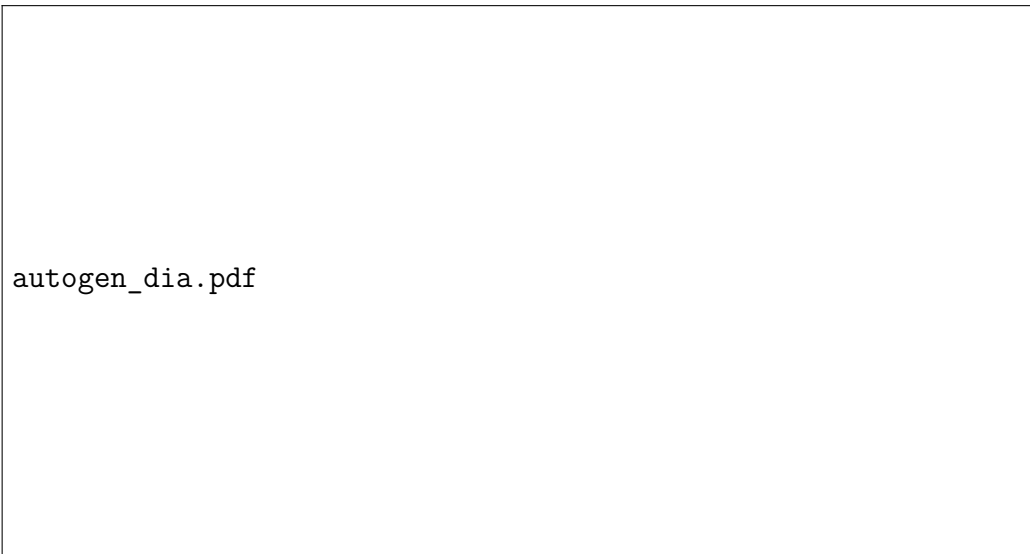


Figure 3: A detailed diagram for automatic conversation generation for two speakers. The first step is to extract clips where only one person is speaking throughout the clip. For each such clip extract dialogues and noise profiles. Shuffle the dialogues and noise profiles to generate Autogen1. For Autogen2, dialogues are embedded with noise profiles in the post and prefix location of the clip. Further, 20 noise profiles are retained from audio clips not present in the dataset. Now, these dialogues and noises are sequenced together to generate conversations.

classified as speech), missed detection (duration of speech classified as non-speech), and confusion (duration of miss classified speaker) divided by the total time duration of the conversation. The formula for DER is:

$$DER = \frac{F.A + M.D + C}{T}$$

where F.A. is a false alarm, M.D is missed detection, C is confusion and T is the total duration of the clip. Traditional approaches to speaker diarization focus on extracting speaker embeddings at frame level using traditional clustering methods [36, 37, 38, 39]. i-vectors, d-vectors, and x-vectors are some commonly used speaker embeddings [40, 41, 42]. Clustering methods are robust for an unidentified number of speakers but suffer optimization for minimizing diarization loss as it is an unsupervised learning method. Furthermore, overlapping speech cannot be determined using these techniques. Cluster-free methods include an end-to-end neural diarization (EEND) which robustly identifies speakers and the overlapped speech but is restricted to a fixed number of speakers [43]. The same authors resolve this limitation by leveraging conditional inference methods based on speaker-wise chain rule [44]. Numerous recent works improve state-of-the-art results by introducing new deep learning architectures for this problem [45, 46, 47, 48, 49, 50]. This problem utilizes rich conversational datasets such as AMI meeting corpus [4], CHIME [6], ICSI meeting corpus [5], and DIHARD challenge datasets [15] for training. Among numerous applications for conversations, CONVURL is one such dataset that can be used for speaker diarization problem targeting Urdu conversations in the wild. To the best of our knowledge, we are the first ones to provide a speaker diarization system for Urdu.

For the aforementioned task, we utilize Pyannote.audio [13], an open-source, robust toolkit that incorporates trainable modules that can be used and optimized collectively for the speaker diarization task. These modules include speaker activity detection (SAD) for extracting speech segments, speaker change detection (SCD) for identifying speaker change followed by speaker embeddings (EMB) to extract distinct speaker identities, and finally the clustering module for grouping embeddings against speaker identity. Figure 4 elaborates this problem in a step-wise manner. Each task can be individually trained and used collectively for speaker diarization problem. The authors reported state-of-the-art results as Diarization Error Rate (DER) for three datasets namely AMI meeting corpus [4], DIHARD [15] and E-TAPE [7] as 29.6, 34.4 and 24.0 respectively. Speaker embeddings task is trained on

VoxCeleb [51] dataset containing more than 7000 unique speakers with one million utterances worth 2000 hours. In the following section, we describe our experiments in detail.

4.2. Results of experimentation

As explained in the previous section, we have utilized Pyannote audio version 1.1.1 ¹ for speaker diarization with the *default settings* for each pipeline [13] on the machine with Nvidia 24 GB Quadro P6000 GPU, Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz with CentOS Linux 7 (Core) operating system. We have used 0-ms collar with overlap while evaluating our results. While we have trained SAD and SCD modules from scratch for 10 epochs respectively, we have used a speaker embedding model trained on the VoxCeleb dataset for the embedding task. We directly apply our dataset to get the speaker embeddings. Following the trend of state-of-the-art datasets [4, 6]; we have pre-defined the splits for the whole dataset: 80% training, 10% validation, and 10% testing splits for each type of dataset. These splits are based on the time duration of the clips, and speakers can overlap within the splits. For CONVURL, scholars, anchors, and politicians can overlap between the sessions because they are called again in the show or the video is divided into two parts. For Autogen, each speaker is randomly selected, so there is a chance of overlap. We have also included results on dataset based on Autogen1, Autogen2 and CONVURL conversations. We call it *Autogen3*. The training split of CONVURL is merged with 50 conversations of Autogen1 and 50 conversations of Autogen2. The validation and testing splits of CONVURL are each merged with 10 conversations of Autogen1 and Autogen2 each. Details against each dataset are shown in Table 3.

For the CONVURL dataset, after training the whole pipeline, we get 29.12 DER(%) with 71.63% correct detection, 0.84% false alarm, 6.86% missed detection and 21.51% confusion respectively. We have a high confusion rate since the Pyannote pipeline does not help in detecting overlap. For automatically generated conversations, we set the number of speakers per conversation n as 5, because we wanted to compare these results with the CONVURL dataset. Further utilizing the algorithm 1 described in the former section, we set $m = 150, t = 40$ and $m = 20, t = 40$ for training and validation/testing splits. For the Autogen1 dataset, after training the

¹<https://github.com/pyannote/pyannote-audio>

whole pipeline and testing on CONVURL, we get 68.91 DER(%) with 31.63% correct detection, 0.54% false alarm, 33.85% missed detection, and 34.51% confusion respectively. For the Autogen2 dataset, after training the whole pipeline and testing on CONVURL, we get 65.32 DER(%) with 37.74% correct detection, 3.06% false alarm, 4.82% missed detection, and 57.44% confusion respectively. We observe that the error is huge because no fine-tuning has been performed at this step. We have applied the CONVURL test set because the test case will always be a real dataset, not a synthetic one. Furthermore, we train the whole pipeline for Autogen3. This variation contains Autogen1+Autogen2+CONVURL conversations in all train, validation, and test splits. We get 28.97 DER(%) with 73.94% correct detection, 2.91% false alarm, 10.51% missed detection, and 15.51% confusion respectively. We have also optimized the pipeline for CONVURL validation split and tested on its test set, we get 30.26 DER(%) with 71.77% correct detection, 2.02% false alarm, 8.90% missed detection, and 19.33% confusion respectively. Next, we fine-tune CONVURL on trained SAD and SCD modules of Autogen1 and Autogen2 respectively for testing the quality of the auto-generated dataset and if it could be used for training speaker diarization pipelines. We get 25.35 DER(%) with 75.21% correct detection, 0.56% false alarm, 11.96% missed detection, and 12.64% confusion respectively when CONVURL fine-tuned on Autogen1. We get 25.26 DER(%) with 75.80% correct detection, 1.06% false alarm, 10.72% missed detection, and 13.48% confusion respectively when CONVURL fine-tuned on Autogen2. As observed, there is no difference between the results, but we get 25% DER for both settings, and the error has been reduced from 29% to 25%. When we directly apply the CONVURL’s test set on the trained Autogen model, we get DER 68.91% and 65.32% for Autogen1 and Autogen2 respectively. This shows the fact that the synthetically generated dataset can be used for training purposes without the need to prepare extensive datasets. For cross-lingual experimentation, we also fine-tuned the AMI meeting corpus on the trained Autogen3 models. We get 43.00 DER(%) with 57.62% correct detection, 0.62% false alarm, 32.56% missed detection, and 9.82% confusion, respectively when AMI fine-tuned on Autogen3. We also fine-tune Autogen3 on the trained AMI models. We get 31.37 DER(%) with 71.03% correct detection, 2.41% false alarm, 8.85% missed detection, and 20.11% confusion, respectively when Autogen3 fine-tuned on AMI. All these results are elaborated in Table 4. From the results, we conclude that both variations are suitable for natural conversations in the Urdu language, especially if finetuned individually on Autogen1 and Auto-

Table 3: Dataset split statistics with Name, training (Tr.), validation (Val.) and testing (Test) splits.

Name	Tr.	Val.	Test
CONVURL	18 h 59 m	2 h 41 m	2 h 34 m
Autogen1	102 h 54 m	13 h 43 m	13 h 28 m
Autogen2	102 h 28 m	13 h 23 m	13 h 22 m
Autogen3	67 h 55 m	16 h 12 m	16 h 27 m


Table 4: Speaker Diarization results on various datasets: CONVURL, Autogen1, Autogen2, Autogen3, and AMI. DER % is reported based on correct detection, false alarm, missed detection, and confusion. Finetune model represents that the dataset in the train, validation, and test set has been fine-tuned on. Results with training and testing on AMI corpus reported in [13]. We add the results shown on their GitHub repository.

Train	Test	Finetune Model	DER%
CONVURL	CONVURL	-	29
Autogen1	CONVURL	-	68
Autogen2	CONVURL	-	65
CONVURL	CONVURL	Autogen1	25
CONVURL	CONVURL	Autogen2	25
Autogen3	Autogen3	-	28
Autogen3	CONVURL	Autogen3	30
Autogen3	Autogen3	AMI	31
AMI	AMI	-	32
AMI	AMI	Autogen3	43

gen2 trained models. However, since the AMI corpus consists of 100 hours of speech, it did not converge well with the baseline settings hence reaching an error rate of 43%. More optimization is required and parameter tweaking of individual models is required to achieve better results. Lastly, fine-tuning is a necessary step for achieving good results in natural conversations.

5. Discussion

The section elaborates the process, implications, and challenges faced during data collection followed by the application of conversations. Lastly, we also highlight the limitations of our work.



speakerdiarization.pdf

Figure 4: A schematic diagram for speaker diarization problem. The module is to perform SAD, followed by SCD. The next step is to extract speaker embeddings and cluster them accordingly. After re-segmentation, predicted labels are assigned to audio chunks.

5.1. First Conversational Dataset for Urdu

In this work, we present the first conversational dataset containing spontaneous speech based on scholarly discussions and political debates for Urdu. This contribution will open a novel avenue for research and also be useful for native Urdu speakers of Pakistan. Speaker identification and validation, active and passive speaker recognition, identifying gender discrimination by analyzing speech turns allowed for a particular speaker involved in a conversation, topic identification and summarization of a conversation are some of the applications which native speakers can utilize to discover interesting insights.

5.2. Scalable approach to generate language-agnostic datasets

Keeping in mind the overhead involved in the video selection and annotation process, we provide a scalable approach to automatically generate datasets for any language with a variable number of speakers and time duration using unique speaker audio profiles. We provide two algorithms that generate such conversations eliminating the need of spending time and resources for procuring a conversational dataset. Furthermore, the experimentation on such datasets for speaker diarization shows that results are comparable with the conversations in the wild (CONVURL) hence, eliminating the time and resource constraint which is typically involved in the creation of all datasets. Furthermore, irrespective of the context, the idea is to mimic real scenario conversations that could be used for training, validation, and testing some applications of conversations. Lastly, such an algorithm would create hours long repository by utilizing just a handful of speaker profiles.

5.3. Real conversation scenarios

The Autogen (algorithm 1) embeds ambient noise within dialogues and mimics the scenarios involved in real-time conversations in the following ways.

Let's consider a scenario where a telephonic or conference call is concerned, each speaker has its ambient noise and variable mic quality and distance. Further, conversations are involved within the same conference room where speakers are seated around the table with the same ambient noise but different microphone distances. These all scenarios are being mimicked by Autogen1 conversations where random silences are embedded within dialogues. Further, there is no restriction on the number of speakers within a conversation and our algorithm provides that functionality. Lastly, Autogen2 handles more robust cases where each speaker's ambient noise is embedded along with silence profiles not present in speakers' clips.

5.4. Applications

Communication among humans mostly takes place in form of conversations. These discussions inherently manifest meaningful information regarding participants' behaviors, levels of interactivity, and intentions. Automatic systems for detecting personality traits have been developed using a simulated tourist call center's conversations [1]. A recent work utilizes the use of capsule neural networks for personality trait detection for speakers [52]. Further applications include automatic empathy recognition from conversations [2], mood and emotion detection from conversations [53], suicide ideation [54], quantifying privacy in human interactions [55] and identification of competitiveness and cooperation in speech overlaps occurring in daily interactions [56]. All above-mentioned applications would prove fruitful if applied to conversations spoken in Urdu.

Overlaps and speaker turns are inherent features for any natural spontaneous human interactions. Consider a meeting/debate/talk show scenario, it would be interesting to discern if all speakers are given the same time to elicit their views, inequality in screen time would indicate either passiveness or bias. The Addition of gender information can help in identifying gender issues. Incorporated with emotional cues, a system could be developed that automatically recognizes the aforementioned traits present in one-to-one human conversations. Speaker diarization problem could be utilized and enhanced for such applications.

5.5. Limitations

This work provides the first speech conversational dataset for Urdu and a universal algorithm for conversation generation with limitations. The fact of employing one annotator for extracting ground truth is our limitation

for this work due to limited resources and we plan to improve this in the future. Regarding the spontaneous speech dataset, the duration of the entire repository is not comparable with the state-of-the-art datasets available for other languages in terms of time duration. We have made this repository available for researchers to add diverse conversations covering various topics.

Further, our automatically generated conversations are void of context because we are embedding different dialogues and silences randomly from different speaker profiles. Moreover, the dialogue of one particular speaker could be segmented in the middle where a break in speech is considered as silence by the algorithm, the conversation generated will not have any semantic meaning. Generating context-dependent conversations is our future work.

Moreover, these conversations do not appear natural when compared with spontaneous human conversations because of incomplete dialogues and abrupt silence embeddings. This problem would be solved if by devising an algorithm that generates conversations on dialogues hence providing a context and a natural flow to a discussion.

The natural conversation includes an overlap in speaker turns where more than one speaker is talking simultaneously. Our algorithm does not generate such conversations. It would be an interesting direction to pursue while an appropriate speaker diarization pipeline handling such overlaps should also be utilized as the current model does not train on overlapped speech.

6. Conclusion

In this study, we have developed the first conversational repository for Urdu. Secondly, keeping in mind the arduous task of speech tagging, we propose an algorithm that produces language-agnostic conversations of variable time and number of speakers in a scalable manner. For evaluation of our dataset, we have chosen the speaker diarization problem and reported DER in various experimental settings. From our results we conclude that these datasets can prove fruitful for training speaker diarization pipelines without the preparation of huge datasets, hence saving a lot of time and effort.

References

- [1] A. V. Ivanov, G. Riccardi, A. J. Sporcka, J. Franc, Recognition of personality traits from human spoken conversations, in: Twelfth Annual Conference of the International Speech Communication Association, 2011.

- [2] F. Alam, M. Danieli, G. Riccardi, Annotating and modeling empathy in spoken conversations, *Comput. Speech Lang.* 50 (C) (2018) 40–61. doi:10.1016/j.csl.2017.12.003. URL <https://doi.org/10.1016/j.csl.2017.12.003>
- [3] M. Danieli, G. Riccardi, F. Alam, Emotion unfolding and affective scenes: A case study in spoken conversations, in: *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies, ERM4CT '15*, Association for Computing Machinery, New York, NY, USA, 2015, p. 5–11. doi:10.1145/2829966.2829967. URL <https://doi.org/10.1145/2829966.2829967>
- [4] J. Carletta, Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus, *Language Resources and Evaluation* 41 (2) (2007) 181–190.
- [5] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Piskin, T. Pfau, E. Shriberg, A. Stolcke, et al., The icsi meeting corpus, in: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 1, IEEE, 2003, pp. I–I.
- [6] J. Barker, S. Watanabe, E. Vincent, J. Trmal, The fifth'chime'speech separation and recognition challenge: dataset, task and baselines, *arXiv preprint arXiv:1803.10609* (2018).
- [7] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, O. Galibert, The ETAPE corpus for the evaluation of speech-based TV content processing in the French language, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 114–118. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/495_Paper.pdf
- [8] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, D. Graff, Hkust/mts: A very large scale mandarin telephone speech corpus, in: Q. Huo, B. Ma, E.-S. Chng, H. Li (Eds.), *Chinese Spoken Language Processing*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 724–735.

- [9] D. M. Eberhard, G. F. Simons, C. D. Fennig, *Ethnologue: Languages of the world*. 23rd edn. Dallas (2020).
- [10] A. A. Raza, A. Athar, S. Randhawa, Z. Tariq, M. B. Saleem, H. Bin Zia, U. Saif, R. Rosenfeld, Rapid collection of spontaneous speech corpora using telephonic community forums, in: *Proc. Interspeech 2018*, 2018, pp. 1021–1025. doi:10.21437/Interspeech.2018-1139.
URL <http://dx.doi.org/10.21437/Interspeech.2018-1139>
- [11] K. Walker, X. Ma, D. Graff, S. Strassel, S. Sessa, K. Jones, RATS Speech Activity Detection (2015). doi:11272.1/AB2/1UISJ7.
URL <https://hdl.handle.net/11272.1/AB2/1UISJ7>
- [12] J. G. Fiscus, J. Ajot, J. S. Garofolo, The rich transcription 2007 meeting recognition evaluation, in: R. Stiefelhagen, R. Bowers, J. Fiscus (Eds.), *Multimodal Technologies for Perception of Humans*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 373–389.
- [13] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, M. Gill, Pyanote.audio: Neural building blocks for speaker diarization, in: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7124–7128. doi:10.1109/ICASSP40776.2020.9052974.
- [14] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, A. Zisserman, Spot the Conversation: Speaker Diarisation in the Wild, in: *Proc. Interspeech 2020*, 2020, pp. 299–303. doi:10.21437/Interspeech.2020-2337.
URL <http://dx.doi.org/10.21437/Interspeech.2020-2337>
- [15] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines, in: *Proc. Interspeech 2019*, 2019, pp. 978–982. doi:10.21437/Interspeech.2019-1268.
URL <http://dx.doi.org/10.21437/Interspeech.2019-1268>
- [16] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, D. House, Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture., in: *LREC*, 2010, pp. 2992–2995.

- [17] B. C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha, S. R. M. Prasanna, Multi-variability speech database for robust speaker recognition, in: 2011 National Conference on Communications (NCC), 2011, pp. 1–5. doi:10.1109/NCC.2011.5734775.
- [18] S. Dalmia, X. Li, F. Metze, A. W. Black, Domain robust feature extraction for rapid low resource asr development, in: 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 258–265. doi:10.1109/SLT.2018.8639569.
- [19] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen, Speech corpus development for a speaker independent spontaneous urdu speech recognition system, Proceedings of the O-COCOSDA, Kathmandu, Nepal (2010).
- [20] L. Shang, Z. Lu, H. Li, Neural responding machine for short-text conversation, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1577–1586.
- [21] Y. Zhuang, X. Wang, H. Zhang, J. Xie, X. Zhu, An ensemble approach to conversation generation, in: National CCF Conference on Natural Language Processing and Chinese Computing, Springer, 2017, pp. 51–62.
- [22] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, D. Yin, Knowledge diffusion for neural dialogue generation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1489–1498.
- [23] W.-N. Zhang, L. Li, D. Cao, T. Liu, Exploring implicit feedback for open domain conversation generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
- [24] J. Xu, H. Wang, Z. Niu, H. Wu, W. Che, Knowledge graph grounded goal planning for open-domain conversation generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 9338–9345.

- [25] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, A. Zisserman, Spot the Conversation: Speaker Diarisation in the Wild, in: Proc. Interspeech 2020, 2020, pp. 299–303. doi:10.21437/Interspeech.2020-2337.
- [26] S. Khan, J. Basu, M. Pal, R. Roy, M. S. Bepari, Multilingual conversational telephony speech corpus creation for real world speaker diarization and recognition, in: 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, 2016, pp. 177–182.
- [27] H. Mao, S. Li, J. McAuley, G. Cottrell, Speech recognition and multi-speaker diarization of long conversations, 2020, pp. 691–695. doi:10.21437/Interspeech.2020-3039.
- [28] C. Cieri, D. Miller, K. Walker, Research methodologies, observations and outcomes in (conversational) speech data collection, in: Proc. HLT 2002, 2002.
- [29] M. S. Grover, P. Bamdev, Y. Kumar, M. Hama, R. R. Shah, audino: A modern annotation tool for audio and speech (2020). arXiv:2006.05236.
- [30] A. Berg, J. Johnander, F. Durand de Gevigney, J. Ahlberg, M. Felberg, Semi-automatic annotation of objects in visual-thermal video, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 2242–2251. doi:10.1109/ICCVW.2019.00277.
- [31] P. Boersma, D. Weenink, Praat: doing phonetics by computer (version 5.1.13) (2009).
URL <http://www.praat.org>
- [32] G. Levy, R. Sitman, I. Amir, E. Golshtein, R. Mochary, E. Reshef, R. Reichart, O. Allouche, GECKO — A Tool for Effective Annotation of Human Conversations, in: Proc. Interspeech 2019, 2019, pp. 3677–3678.
- [33] H. Titeux, R. Riad, pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in python, Journal of Open Source Software 6 (62) (2021) 2989. doi:10.21105/joss.02989.
URL <https://doi.org/10.21105/joss.02989>

- [34] J. Robert, M. Webbie, et al., Pydub (2018).
URL <http://pydub.com/>
- [35] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, S. Narayanan, A review of speaker diarization: Recent advances with deep learning (2021). arXiv:2101.09624.
- [36] S. H. Shum, N. Dehak, R. Dehak, J. R. Glass, Unsupervised methods for speaker diarization: An integrated and iterative approach, *IEEE Transactions on Audio, Speech, and Language Processing* 21 (10) (2013) 2015–2028.
- [37] G. Sell, D. Garcia-Romero, Speaker diarization with plda i-vector scoring and unsupervised calibration, in: 2014 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2014, pp. 413–417.
- [38] D. Dimitriadis, P. Fousek, Developing on-line speaker diarization system., in: INTERSPEECH, 2017, pp. 2739–2743.
- [39] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, I. L. Moreno, Speaker diarization with lstm, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5239–5243.
- [40] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing* 19 (4) (2010) 788–798.
- [41] L. Wan, Q. Wang, A. Papir, I. L. Moreno, Generalized end-to-end loss for speaker verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4879–4883.
- [42] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust dnn embeddings for speaker recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5329–5333.
- [43] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with self-attention, in: 2019 IEEE

Automatic Speech Recognition and Understanding Workshop (ASRU),
IEEE, 2019, pp. 296–303.

- [44] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, J. Shi, K. Nagamatsu, Neural speaker diarization with speaker-wise chain rule, arXiv preprint arXiv:2006.01796 (2020).
- [45] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, K. Nagamatsu, End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors, in: Proc. Interspeech 2020, 2020, pp. 269–273. doi:10.21437/Interspeech.2020-1022.
URL <http://dx.doi.org/10.21437/Interspeech.2020-1022>
- [46] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko, Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario, in: Proc. Interspeech 2020, 2020, pp. 274–278. doi:10.21437/Interspeech.2020-1602.
URL <http://dx.doi.org/10.21437/Interspeech.2020-1602>
- [47] Q. Lin, Y. Hou, M. Li, Self-Attentive Similarity Measurement Strategies in Speaker Diarization, in: Proc. Interspeech 2020, 2020, pp. 284–288. doi:10.21437/Interspeech.2020-1908.
URL <http://dx.doi.org/10.21437/Interspeech.2020-1908>
- [48] N. Dawalatabad, S. Madikeri, C. C. Sekhar, H. A. Murthy, Novel architectures for unsupervised information bottleneck based speaker diarization of meetings, IEEE/ACM Trans. Audio, Speech and Lang. Proc. 29 (2021) 14–27. doi:10.1109/TASLP.2020.3036231.
URL <https://doi.org/10.1109/TASLP.2020.3036231>
- [49] P. Singh, S. Ganapathy, Self-supervised representation learning with path integral clustering for speaker diarization, IEEE/ACM Trans. Audio, Speech and Lang. Proc. 29 (2021) 1639–1649. doi:10.1109/TASLP.2021.3075100.
URL <https://doi.org/10.1109/TASLP.2021.3075100>
- [50] M. Pal, M. Kumar, R. Peri, T. J. Park, S. H. Kim, C. Lord, S. Bishop, S. Narayanan, Meta-learning with latent space clustering in generative adversarial network for speaker diarization,

- IEEE/ACM Trans. Audio, Speech and Lang. Proc. 29 (2021) 1204–1219.
doi:10.1109/TASLP.2021.3061885.
URL <https://doi.org/10.1109/TASLP.2021.3061885>
- [51] J. S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, in: Proc. Interspeech 2018, 2018, pp. 1086–1090.
doi:10.21437/Interspeech.2018-1929.
URL <http://dx.doi.org/10.21437/Interspeech.2018-1929>
- [52] E. A. Rissola, S. A. Bahrainian, F. Crestani, Personality recognition in conversations using capsule neural networks, in: IEEE/WIC/ACM International Conference on Web Intelligence, WI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 180–187.
doi:10.1145/3350546.3352516.
URL <https://doi.org/10.1145/3350546.3352516>
- [53] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, E. Mower Provost, The priori emotion dataset: Linking mood to emotion detected in-the-wild, in: Proc. Interspeech 2018, 2018, pp. 1903–1907. doi:10.21437/Interspeech.2018-2355.
URL <http://dx.doi.org/10.21437/Interspeech.2018-2355>
- [54] J. Gideon, H. T. Schatten, M. G. McInnis, E. M. Provost, Emotion Recognition from Natural Phone Conversations in Individuals with and without Recent Suicidal Ideation, in: Proc. Interspeech 2019, 2019, pp. 3282–3286. doi:10.21437/Interspeech.2019-1830.
URL <http://dx.doi.org/10.21437/Interspeech.2019-1830>
- [55] P. P. Zarazaga, S. Das, T. Bäckström, V. V. R. V., A. K. Vuppala, Sound Privacy: A Conversational Speech Corpus for Quantifying the Experience of Privacy, in: Proc. Interspeech 2019, 2019, pp. 3720–3724. doi:10.21437/Interspeech.2019-1172.
URL <http://dx.doi.org/10.21437/Interspeech.2019-1172>
- [56] K. Truong, Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee, in: Proceedings of the 14th Annual Conference of the International Speech Communication Association, Interspeech 2013, International Speech Communication Association (ISCA), 2013, pp. 1404–1408.
URL <http://www.interspeech2013.org/>